DOCUMENT RESUME

ED 441 849 TM 030 884

AUTHOR	Whetton, Chris; Twist, Elizabeth; Sainsbury, Marian
TITLE	National Tests and Target Setting: Maintaining Consistent Standards.
INSTITUTION	National Foundation for Educational Research, Slough (England).
PUB DATE	2000-04-00
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
AVAILABLE FROM	For full text: http://www.nfer.ac.uk.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Standards; Accountability; Educational Trends; *Equated Scores; Foreign Countries; *National Competency Tests; National Curriculum; Scoring; *Test Construction
IDENTIFIERS	Angoff Methods; *England; Scripts (Knowledge Structures); *Standard Setting; Target Planning

ABSTRACT

It is now seen as an economic and political necessity for countries to produce higher levels of performance across the spectrum of ability of all their students. This paper describes one example of the influence of political conditions on the process of developing assessment instruments and on measuring standards. In England, the trend is toward total accountability in education, from the National Curriculum introduced in 1988 to the National Curriculum assessment system that has been implemented since about 1996. The widespread use of target setting as a management and motivational device has become part of the assessment and curriculum process. At the start of the current contracts for testing, covering tests for the years 2000 to 2002, the government's Qualifications and Curriculum Authority and the National Foundation for Educational Research agreed that no one method of standard setting could stand public scrutiny. For this reason, four methods are being used to assure the constant nature of standards: (1) direct statistical equating with the test from the previous year; (2) equating of the new test with an anchor test; (3) an Angoff-type standard setting procedure; and (4) a "script scrutiny" procedure. Each of these approaches is described. These processes illustrate that the setting and maintenance of standards is a social and societal process that can stand only if it is acceptable publicly and politically. (Contains 20 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 849

National Tests and Target Setting: Maintaining Consistent Standards

Chris Whetton, Elizabeth Twist and Marian Sainsbury

c.whetton@nfer.ac.uk
l.twist@nfer.ac.uk
m.sainsbury@nfer.ac.uk

National Foundation for Educational Research

The Mere, Upton Park, Slough, Berkshire SL1 2DQ

UNITED KINGDOM

www.nfer.ac.uk

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

C. Whetton

1

TM030884

Paper presented at American Educational Research Association

Annual Meeting 2000 in New Orleans

BEST COPY AVAILABLE

National Tests and Target Setting: Monitoring Consistent Standards

Background

All around the world, education systems are undergoing processes of change. These may involve changes to the curriculum, refinements to assessment systems, alterations in management structures and responsibilities, the introduction of inspection or other accountability mechanisms and shifts in control towards or away from schools. (OECD, 1995; Stevenson, 1996) In many countries, the motivation for such change is political and economic. Often, underlying the desire for change is a perceived need to raise educational standards because these are believed to have a link to future economic success. For some, this link is direct and is supported by international surveys which show the highest performing countries educationally are often those with economies which have been very successful. For others, the link is more indirect but does acknowledge that the influence of modern communications in creating world-wide markets for goods and services, effectively means that countries are competing economically and will require highly educated workforces derived from better education systems (Brown and Lauder, 1996). In short, it is now seen as an economic and political necessity for countries to produce higher levels of performance across the spectrum of ability of all their students.

This paper will set out one example of the influence of these political conditions on the process of developing assessment instruments and on measuring standards.

The influence of assessment systems in introducing change in education is well known (see for example Keeves, 1994). When the stakes are high and the results matter, the tests provide many types of influence. They influence what is taught, how it is taught, the motivation of the teachers and the motivation of the students. They can be used to reinforce aspects of the curriculum or to signal that others are less important. For all these reasons, policy makers, opinion formers and politicians believe that changes to assessment systems can bring about change to the larger education system and, generally, they wish to use them to raise standards. An aspect of this process which has been less commented upon is that the assessment systems, as well as being the instruments of change, may also provide the measures of that

change. This differs from other areas of social policy, where a policy and its implementation are usually distinct from the measure of its success. For example the success of changes to food hygiene regulations would be measured by a completely separate outcome, the incidence of food poisoning. The dual role of assessment systems as the instruments of reform and also the measures of change makes the procedures used open to scrutiny from two perspectives, and takes them into the arena of public debate both about their effectiveness and also their reliability and validity as measures.

On the spectrum of accountability in education, the United Kingdom, and particularly England is towards one extreme: that of total accountability. A National Curriculum was introduced for the first time in 1988. The original blueprint was set out in a working party report (GB. DES and WO, 1988). Previously schools had responsibility for their own curricula, with guidance provided by local education authorities. This National Curriculum set out what children should be taught from the ages of 5 to 16 and also gave a series of expectations for what they should be able to do at various ages. These expectations were termed "attainment targets". Alongside the curriculum, compulsory testing of all pupils in state schools was introduced for students at the ages of seven, 11 and 14, adding to the school leaving examinations already in place for 16-year-olds. These tests were originally known as standard assessment tasks (SATs) but since 1991, have been termed National Curriculum Tests. The tests are in mathematics, in English (reading and writing only) and, except for the seven-year-olds, in science. It should be reiterated that this is not a sample-based monitoring system but is the compulsory testing of the complete population of children, which is about 600,000 in each age group. Over the decade, the assessment system changed many times, settling down in about 1996 (Shorrocks-Taylor, 1999). International comparisons tend to show that England is the country with the most testing for the most purposes, certainly among developed Western nations (eg. Whetton, 1999).

The curriculum and testing regime was only a part of a general upheaval in the education system. Financial management of their budgets was devolved to the schools. Hence while the curriculum was centralised, financial management was de-centralised. Further accountability measures were also introduced. Regular

inspections of schools were begun, with failing schools subject to special measures and close monitoring with the ultimate threat of closure.

The model behind these changes was one of competition between schools in which pressure from parents would raise standards. Hence, the reports of the school inspections are made public and the test results for 11-year-olds (the end of primary school) and 16-year-olds (the end of compulsory schooling) are published on a school-by-school basis. This is then a high-stakes assessment system in which the stakes are greatest for teachers and school managers. For 11-year-olds the consequences for individual pupils are minor. Formally, their progression in the education system does not depend on the results. There is nevertheless a great deal of parental and school pressure on the pupils.

From the start, the unfairness of the publication of these unadjusted results has been trumpeted by those working in areas of social deprivation, such as inner cities and there have been efforts to develop 'value added' measures which show progress or performance adjusted for previous attainment or social background (Saunders, 1999).

The National Curriculum assessment system was originally intended to be criterion-referenced with the attainment targets set out as levels of attainment. As time has passed, strict criterion-referencing has been replaced by a looser criterion-related system (Sainsbury and Sizmur, 1998). The levels are defined by a description which sets out what those achieving it can do. Two examples are shown in Figure 1: for level 3 and level 4 of the Reading attainment target. These level descriptions are necessarily very broad and the original conception was that an increase of one level would reflect two years of teaching and learning. Since schooling starts at age 5, this was level 1, level 2 corresponded to the average seven-year-old and so on. The levels were not derived empirically but were expectations of what children should be able to do at various ages. Hence level 4 was the expected achievement of an average 11-year-old.

Figure 1: Examples of level descriptions of attainment target: Reading, levels 3 and 4

Level 3

Pupils read a range of texts fluently and accurately. They read independently, using strategies appropriately to establish meaning. In responding to fiction and non-fiction they show understanding of the main points and express preferences. They use their knowledge of the alphabet to locate books and find information.

Level 4

In responding to a range of texts, pupils show understanding of significant ideas, themes, events and characters, beginning to use inference and deduction. They refer to the text when explaining their views. They locate and use ideas and information.

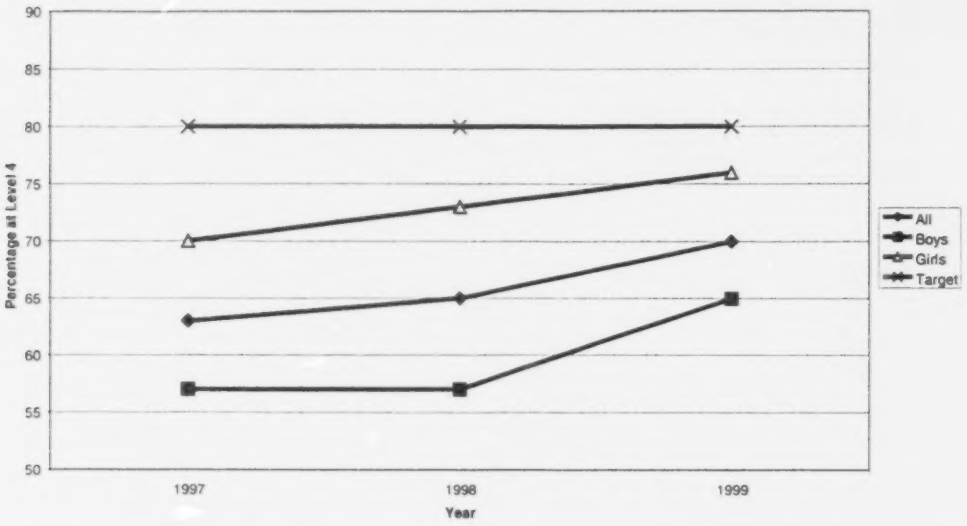
The published tables of results for schools focus on level 4, giving the percentage of students who have achieved that level or higher in English, mathematics and science. Hence, the most vital part of the system is the cut-score for the achievement of level 4. The results are published in national newspapers and are available locally and on an internet site.

In 1997, a new Labour Government was elected and immediately placed education at the head of its priorities. Rather than abolishing any of the existing accountability measures inherited from the previous Conservative government, it built further on them. In particular, it emphasised the importance of higher achievement in literacy. A compulsory scheme, the National Literacy Strategy (GB. DfEE, 1997), was introduced which required all primary schools to devote an hour each day for every child to teaching literacy. The content and even the pedagogic methods were prescribed. This went beyond the original National Curriculum requirements, which had specifically allowed teachers to use their own methods. A similar strategy has also been introduced for numeracy and both are being extended into secondary schools.

A second new aspect introduced was the widespread use of target setting as a management and motivational device. Schools are required to set targets for individual pupils stating what they are to learn next. They must also set targets for the

school as a whole. One of these targets has to be in terms of the proportion of pupils who will achieve level 4 in the National Curriculum tests in English. To provide a lead in this process, the Government set an overall target that 80 per cent of 11-year-old children should achieve level 4 or higher in English by the year 2002, the latest end of the current parliament. In 1997, when the target was set, the figure was 63 per cent, so an increase of 17 per cent over five years was required. The politician responsible, the Secretary of State for Education and Employment, David Blunkett, promised to resign if the target was not met. Hence, the test results became high stakes for politicians and their opponents, as well as schools, teachers and students. Progress so far towards this target is shown in Figure 2.

Figure 2: Percentage of Pupils Achieving Level 4



This has led to greater scrutiny of the tests, test development methods and most of all the standard setting process. For reasons of security and also to have a constantly changing curriculum backwash effect, the tests are changed each year. The problem of maintaining the standard of level 4 as a constant becomes a major one for the test developers. The problem is both technical and one of public perception. Attempting to meet its requirements poses fundamental questions about the nature of standards and the means by which a standard can be kept constant.

The development of the National Curriculum tests in English for 11-year-olds is undertaken by a contractor, the National Foundation for Educational Research

(NFER). The government agency responsible is the Qualifications and Curriculum Authority (QCA).

At the start of the current contracts which covers the tests for the years 2000 to 2002, both NFER and QCA were aware of the importance of the standards setting procedure. It was agreed that no one method alone could stand public scrutiny. Hence, a variety of different methods were, and are, utilised, but each itself has both disadvantages and advantages. There is also the further problem of combining information from several sources to arrive at a single decision on the cut score.

The four methods used are: direct statistical equating of the new test with that of the previous year; equating of the new test with an anchor test which has been used for several years; an Angoff-type standard setting meeting; and a 'script scrutiny' procedure. Hence, two of the methods are based on empirical data and two are based on expert judgement. Descriptions of the procedures used appear in Whetton *et al.* (1999) and Rose (1999).

Statistical equating to previous year's test

Statistical equating from year to year is made possible by the timing of the development schedule. Work begins on each test just over two years before it is to be used. There is a year of early development, including field trials leading to item selection, and the final test is ready a full year before it is needed. This allows a further large-scale trial, known as the final pre-test. For the 1999 test, the final pre-test took place in April 1998, and involved a nationally representative sample of 1337 students in year 6 (aged 10 or 11). In May 1998, these students, together with all year 6 students throughout England and Wales, took their real 1998 key stage test. Scores on this test were collected for the pre-test sample. Essentially this allows the equating of the two tests using a common sample of pupils and then equipercentile equating (see for example, Kolen and Brennan, 1995).

The total test score is made up of 50 per cent for reading and 50 per cent for writing. The maintenance of standards in these two elements is different, because the nature of the scoring system differs. The final decision is a cut-score for English overall (including both reading and writing) but this is calculated by considering reading and writing separately, and then adding the cut-scores together to give totals out of 100.

Reading is assessed using a 45-minute test. This has a separate book of stimulus material which includes several different text types. These might be a short story, a poem, factual material, an entry from an encyclopaedia or any appropriate authentic text. The questions are almost all short answer, and give rise to one, two or three score points. The writing test consists of a 45-minute composition, on a narrative or non-narrative topic, selected from a choice of four stimuli. There is also a spelling test and a handwriting test. The composition is scored according to a criterion-referenced mark scheme, as is the handwriting test. These criterion-referenced mark schemes are essentially the same every year. The difficulty of the spelling test is maintained at the same level from year to year. Since this writing element is directly referenced to the criteria expressed in the national curriculum levels, therefore, statistical equating is not considered appropriate. Cut-scores remain the same every year.

As a matter of interest, however, writing scores are equated each year. This analysis reveals a finding that has become known as the 'pre-test effect'. For the same students, for a task assessed by the same criteria, the mean writing scores in the pre-test are around two marks out of 50 lower than scores in the live test (see Table 1). There is some evidence that the size of the effect is increasing. This seems to be due to a difference in motivation, with both students and their teachers highly motivated for the high-stakes live test, and less so for the pre-test, which they know is low-stakes. This pre-test effect is likely also to apply to the reading test, but it is more difficult to quantify, as there is a different test every year. The results are shown in Table 2. Again a clear effect can be seen which may be increasing. It is one reason why caution should be applied to equating results, reinforcing the desirability of drawing on a variety of approaches in determining cut-scores.

Table 1: Size of Pre-test Effect in Writing Test

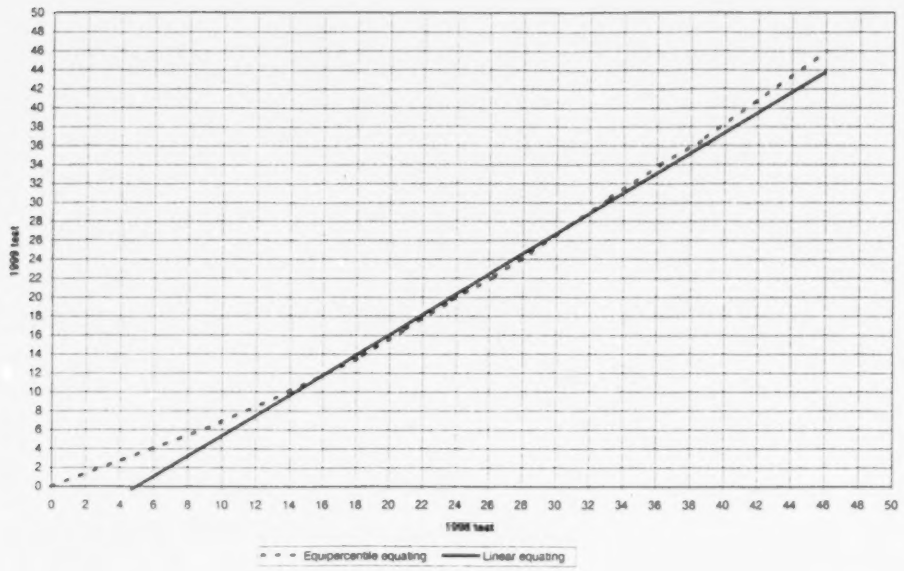
Year	Test Used in Pre-test		Test Used as National Curriculum Assessment	Gain	Gain as % Sd
	Mean	Sd	Mean		
1999	28.9	7.4	31.6	2.7	36
1998	29.1	7.2	31.0	1.9	26
1997	30.4	7.6	31.6	1.2	16

Table 2: Size of Pre-Test Effect in Reading Test

Year	Test Used in Pre-test		Test Used as National Curriculum Assessment	Gain	Gain as % Sd
	Mean	Sd	Mean		
1999	22.5	9.1	25.7	3.2	35
1998	24.5	8.8	26.1	1.6	18
1997	25.0	10.3	26.9	1.9	18

Scores on the 1998 live reading test were directly equated with the 1999 pre-test. Figure 3 shows the equipercentile equating graph for these two tests.

Figure 3: Equipercentile equating graph of reading scores for the 1998 live test and the 1999 final pre-test



Equating to an anchor test

A further approach to statistical equating is provided by the use of an anchor test. This test has been used every year since 1996. A subsample of around 400 pupils from the final pre-test sample took the 1999 reading test and the anchor test. Scores on the anchor test were equated with those on the 1998 live reading test. Anchor test equivalent cut-scores for every year up to and including 1998 were collated, and a median value taken. This median value was equated with the 1999 pre-test, to give a further set of recommended cut-scores. Again, equipercentile equating is utilised.

There are a number of reasons why the anchor test may be regarded as less reliable than the ideal. There have been subtle changes since 1996 to the way the curriculum is represented in the reading tests. Extended response items yielding three score points are a usual feature of current tests, but none appear in the anchor test. The anchor test sample is smaller than the sample for direct equating. Finally, it has not been possible to keep the timing and nature of the sample for the anchor test the same each year since 1996. In that year, there was an autumn and a February pre-test,

BEST COPY AVAILABLE

rather than the current April pre-test. That made it necessary to include some older students in the pre-testing.

Thus the statistical equating resulting from the final pre-test gives rise to two separate sets of equated cut-scores for the reading test: those arising from direct equating and those arising from equating via the anchor test. These can be added to the unvarying writing cut-scores to give two sets of possible cut-scores for English overall. Both sets of equated cut-scores form part of the final decision-making process. However, each has technical limitations such that they cannot provide a definitive standard.

Angoff Standard Setting Procedure

Two procedures are applied each year in which the opinions of expert judges as to the relative demand of the reading test are obtained. The first of these is a version of the Angoff procedure (Angoff, 1971; Livingstone and Zieky, 1982).

A group of approximately 12 highly experienced teachers are invited to take part in a two-day meeting during which they study the reading test to be taken by all 11-year olds. About half the group have usually been involved in the procedure previously and the remainder will be new to the exercise. The teachers are from a wide variety of schools and regions and are recognised for their expertise in the teaching of English and in particular, their familiarity with the performance of 11-year old students. They are also familiar with the testing process and the context in which the test will be administered nationally later in the year.

The first task of the group is to study the new reading test and the scoring guide. Once they have become familiar with the new materials, they move on to a discussion of the characteristic reading behaviours of students at one particular level of attainment. The National Curriculum means that broad descriptions of attainment are familiar to all teachers. These level descriptions have to be 'fleshed out' into more detailed descriptions of reading behaviours. In the first instance, the judges focus on the level 4 boundary, although subsequently the boundary at the entry level to the test (level 3) and that at level 5 are discussed. This aspect of the meeting is critical as judges refine their notion of the borderline student.

At this point, the judges are trained in the process of recording their individual decisions for each item. Training items are selected which exemplify both single and multiple point questions. For a one point item, judges record the percentage of students, minimally competent at level 4, they expect to be successful on that particular item. For multiple point items, judges have to estimate the percentage of students they would expect to attain one, two or three points. After each training item, the decision of each judge is recorded and displayed, leading to a discussion within the group of the variations in initial judgements. As soon as all judges are clear about the decision-making process they record their estimates on all items without further discussion.

Once the preliminary decisions of all judges have been made, some analyses are undertaken:

for each judge, their estimates as to the likely success rate of a student at a given level boundary on each item are aggregated to give a cut-score;

the decisions of all the judges on each item are aggregated and a mean score derived for each item at each level boundary.

This information is then made available to all judges. Initially they are able to compare their aggregate score (cut-score) with those of their fellow judges at each of the three level boundaries. From this data, they are able to compare their estimate of the demand of the tests as a whole with the estimates of all other judges.

The next piece of statistical information the judges are given is derived from the trial of materials with students in school. Participants are provided with details of the performance of students, grouped according to the level they attained on the previous year's national test taken at the time of the final trial, on each item. This is presented graphically. It is made clear to the judges that the data from the trial describes the mean performance of students across the range of a level and not just at the boundary. They are then provided with the equivalent data derived from the mean estimates of all judges in the group for each item, again graphically.

Judges are then encouraged to discuss as a group the items where the variance is greatest and to consider the demand of those particular items. It is emphasised to

judges that this data is provided for information only and that they should exercise their professional judgement as they continue with the process. No attempt is made to achieve a consensus.

Following a review of the reading behaviours previously identified as being characteristic of students at the particular boundary, judges are then given the opportunity to revise, individually, their initial judgements. When they have completed these they then discuss with two or three other judges items which they have previously identified as significant, often those with the greatest variance.

At the end of the group discussions, participants work alone to finalise their decisions. It is these judgements which are aggregated, after the meeting, and then a mean score for the group overall is calculated at each level boundary. These outcomes are taken to be the recommended cut-scores from the Angoff procedure.

The advantages of the Angoff procedure are that it follows a set procedure of due process and that it gives equal weight to the judges. However, the judges themselves represent only one interest group, the teachers, and the process is very dependent on their knowledge of children of this age. Although they are given statistical information, they do not have access to children's work, showing how they actually answered the questions.

Script Scrutiny

The second element of the standard setting procedure involving the decisions of expert judges is the script scrutiny. This is the only part of the procedure which uses 'live' scripts and takes place after the test has been used throughout England. Representatives of the test development agency attend the meetings as observers but play no part in the proceedings.

National tests are marked independently of the schools by over 1500 markers who are trained each year on the new reading test. Markers are trained using a cascade model and the judges invited to the script scrutiny are between eight and ten of the most senior and experienced of these markers.

Prior to the script scrutiny meeting, the judges are sent nine sets of reading test scripts derived from the tests of the previous three years, which have total scores at exactly one of the three level thresholds agreed for that particular test. This preliminary work enables judges to familiarise themselves with the level of demand of previous tests at each level boundary and to identify features in students' responses which characterise performance at each level threshold.

In advance of the script scrutiny, the agency responsible for the marking of all the tests collates sets of live scripts, each set comprising five scripts with the same total score. These scores are the draft level threshold, and two marks either side of this draft threshold (see below).

The first element of the script scrutiny is similar to the discussions at the start of the Angoff procedure when the characteristics of performance at each of the level thresholds are identified by the participants. In addition to the performance characteristics, at the script scrutiny judges identify specific questions where they feel evidence of performance typical of a particular level is most likely to be found.

Following discussion, for each level boundary, judges independently review the sets of scripts. These scripts show the points awarded for each item but not the total point score. Packs are identified by letter only, allocated randomly.

Judges are charged with the task of identifying which one, if any, of the five packs contains scripts of equivalent performance to the sets of scripts at a particular threshold from the previous tests. The decisions of individual judges as to the performance evident in scripts in a particular set (equivalent to, poorer than or better than performance at the threshold in previous years) are collected and discussed. In contrast to the Angoff procedure, in the script scrutiny an attempt is made to achieve a consensus. If judges agree that one particular pack shows a level of performance equivalent to that at the particular level threshold in previous years, then the total point score of scripts in this set is taken as the recommendation of the script scrutiny panel for the threshold. If a consensus cannot be reached on the packs tabled, then the process continues with either further discussion or the provision of more scripts until the panel feels that it is in a position to make a recommendation. The judges may request to see more scripts at a particular total point band or they may ask for scripts

with a total point score outside the range previously considered. The process continues until the judges arrive at a final recommendation.

The script scrutiny exercise has the advantage that the judges have real scripts from the actual test and hence from children with the correct levels of motivation and curriculum experience. However, the procedures followed do not have the formality of the Angoff meeting and the judgements made may include unintentional biases. Cresswell (1996) who has made an extensive study of script scrutiny methods in English public examinations for 16- and 18-year-olds argues that such methods can work well if awarders share tacit standards, based upon guild knowledge, which are shared with the wider group of examination users. It is not clear if such tacit knowledge can operate when an education system is in change as is the case in England.

Draft Level Thresholds and Level Confirmation

For these National Curriculum tests, marking is undertaken by external markers, not by the children's own teachers. As part of the process, to be fair to children and schools, any pupil who is close to achieving a particular level, but does not quite reach the cut-score has their script re-marked. This means that a draft threshold must be established before the marking takes place, and this utilises only the first three types of evidence listed above.

These draft thresholds or cut-scores are set at a meeting between the test development agency and the responsible government agency (QCA). The meeting considers the equating evidence (both direct and for the Anchor test) and the results of the Angoff meeting. Each person present is asked to weigh these and reach a judgement of the recommended cut-score. Some members choose to emphasise the Angoff evidence, others the equating evidence. The size of the pre-test effect and its effect on the equating are also matters of individual judgement. All the judgements are collated and presented to the meeting before a decision is reached.

About one month after the use of the tests in schools, a 'level confirmation meeting' is held. This again involves the test development agency and the QCA but also contains a number of observers. These include independent academics, representatives of the government department for Education, nominees from teachers and headteacher's

professional associations. This meeting reconsiders the draft thresholds, re-examining the evidence of the equating exercises and the Angoff meeting. However, it also has the further information of the recommendation from the script scrutiny and an estimate of the national distribution of scores, based on a representative sample of 30,000 scripts. This meeting too takes the form of a committee discussion before moving to an agreed decision.

These processes are effectively similar to many committee decisions. Group dynamics operate and some participants are more powerful or more voluble than others. Discussion can focus on relevant issues or introduce extraneous information. These points are made to emphasise that even where there has been a due process or there is apparently hard evidence from statistical equating, decisions reached remain matters of judgement and the standard setting process incorporates these. The legitimacy of the standards resides both in the information used and the process by which it is combined. If that process is regarded as flawed or biased then since there is no objective truth or standard, no possibility of a standard remains.

In 1999, a national newspaper reported that the cut scores for level 4 had been lowered. (Clare, 1999) These were in fact the draft thresholds agreed at the first meeting. The reason for the lower cut-scores was that the 1999 test was harder than that of 1998. This can be seen as an inevitable consequence of one part of the specification, that the mean score should be about 50 per cent of the maximum possible. In an environment in which standards are steadily rising, and the tests constructed to be more difficult, it is inevitable that the constant of a threshold for level 4 should have a lower cut-score. However, rather than accept this technical explanation, the newspaper reported that the cut-score had been lowered in order to assist the Secretary of State to reach the literacy target. To rebut this charge, he set up a committee of enquiry, asking his own and the other main political parties to supply a representative. He also included a *Times* journalist as a member. The committee's chairman was the Deputy Head of the Inspectorate of Schools, Jim Rose.

The committee of enquiry examined all the processes described above and took evidence from academics, test developers and various interest groups as well as teachers and headteachers. They concluded that the concerns about the setting and maintaining of standards of the English test were without foundation. They could find

no evidence whatsoever that the Ministers or officials at the Department for Education and Employment sought to influence the tests in order to meet the national targets in English (Rose, 1999).

Conclusions

In many ways the results of the enquiry were an endorsement of the procedures used, and described in this paper. However, it is legitimate to ask if improvements can be made. A personal view is that a greater element of formality or due process should be present at both the meeting for setting draft thresholds and the confirmation meeting. This is echoing recommendations from US practice (Cizek, 1993 and 1996). For example, it should be made clear who in the meetings is entitled to contribute to the decision and who ultimately is accountable for it. The combination of evidence should be more formal, drawing on Angoff-type procedures. For example, the possibility of agreeing formal weightings for the four types of evidence could be considered. This would allow arithmetic procedures to be used for their combination, rather than private individual judgements. The weightings would be explicit and based on a view of the reliability and importance of the elements. It should be noted, however, that again these are matters of judgement.

The processes outlined here illustrate that the setting and maintenance of standards is a social and societal process. This conclusion is supporting that of Gipps (1999) in the wider sphere that assessment is a social activity that we can only understand by taking account of the social, cultural, economic and political contests in which it operates. The setting of standards does rely on empirical information but that information must be interpreted by those involved. Standard setting relies also on society's willingness to accept the integrity and expertness of those making judgements. If this breaks down, as in some other areas of public policy, then little remains and there is no possibility of maintaining standards over time. It is taken as axiomatic that there needs to be a due process to enhance confidence. However, this alone, though, is not sufficient. The process as a whole from the test construction, through the statistical methods, to the judgemental methods and then to their resolution combines to make standard setting a social process which can only stand if it is acceptable publicly and politically.

References

ANGOFF, W. H. (1971). 'Scales, norms, and equivalent scores.' In: THORNDIKE, R. L. (Ed) *Educational Measurement*. Second edn. Washington, DC: American Council on Education.

BROWN, P. and LAUDER, H. (1996). 'Education, globalization and economic development', *Journal of Education Policy*, **11**, 1, 1-25.

CIZEK, G.J. (1993). 'Reconsidering standards and criteria', *Journal of Educational Measurement*, **30**, 2, 93-106.

CIZEK, G.J. (1996). 'Standard-setting guidelines', *Educational Measurement: Issues and Practice*, **15**, 1, 13-21, 12.

CLARE, J. (1999). 'Passing off', *Daily Telegraph*, 29 May, 29.

CRESSWELL, M. J. (1996). 'Defining, setting and maintaining standards in curriculum-embedded examinations: judgemental and statistical approaches.' In: GOLDSTEIN, H. and LEWIS, T. (Eds) *Assessment: Problems, Developments and Statistical Issues. A Volume of Expert Contributions*. Chichester: John Wiley & Sons.

GIPPS, C. (1999). 'Socio-cultural aspects of assessment.' In: IRAN-NEJAD, A. and PEARSON, P. D. (Eds) *Review of Research in Education 24*. Washington, DC: AERA.

GREAT BRITAIN. DEPARTMENT FOR EDUCATION AND EMPLOYMENT. LITERACY TASK FORCE (1997). *The Implementation of the National Literacy Strategy*. London: DfEE.

GREAT BRITAIN. DEPARTMENT OF EDUCATION AND SCIENCE and WELSH OFFICE (1988). *National Curriculum: Task Group on Assessment and Testing: a Report*. London: DES.

KEEVES, J.P. (1994). *National Examinations: Design, Procedures and Reporting* (Fundamentals of Educational Planning No.50). Paris: UNESCO.

KOLEN, M.J. and BRENNAN, R.L. (1995). *Test Equating: Methods and Practices*. New York, NY: Springer Verlag.

LIVINGSTONE, S.A. and ZIEKY, M.J. (1982). *Passing Scores: a Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: Educational Testing Service.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (1995). *Performance Standards in Education: In Search of Quality*. Paris: OECD.

ROSE, J. (1999). *Weighing the Baby: the Report of the Independent Scrutiny Panel on the 1999 Key Stage 2 National Curriculum Tests in English and Mathematics*. London: DfEE.

SAINSBURY, M. and SIZMUR, S. (1998). 'Level descriptions in the National

Curriculum: what kind of criterion referencing is this?' *Oxford Review of Education*, **24**, 2, 181-93.

SAUNDERS, L. (1999). *'Value Added' Measurement of School Effectiveness: a Critical Review.* Slough: NFER.

SHORROCKS-TAYLOR, D. (1999). *National Testing: Past, Present and Future* (Issues in Assessment and Testing). Leicester: BPS Books.

STEVENSON, D. (1996). 'International patterns of assessment: policy change.' Paper presented at the Annual Meeting of AERA, New York, 8-12 April.

WHETTON, C. (1999). 'Attempting to find the true cost of assessment systems.' Paper presented at the 25th Annual Conference of the International Association for Educational Assessment, Bled, Slovenia, 25 May.

WHETTON, C., SAINSBURY, M. and TWIST, L. (1999). 'Testing times', *Managing Schools Today*, **9**, 3, 14-15.